An Auditable Search Agent Methodology at Scale for Political Elite Research

Final author list TBD

Abstract

Large-N political-elite datasets increasingly rely on digital traces, but scaling elite attribute coding with search-enabled LLMs can undermine auditability, cross-national comparability, and temporal validity. We propose an *auditable search agent* (ASA) protocol that treats web retrieval as a governed measurement instrument. For each $\langle$person, country, year$\rangle$ record, ASA runs a short search session, constrains outputs to expert closed codebooks, enforces an "as-of" evidence rule, abstains under weak or conflicting support, and archives complete provenance (queries, snippets, URLs, timestamps). In a party-affiliation labeling task spanning 114 countries (N = 34,618 leader–records; 1,209 labels), ASA attains 0.860 accuracy on high-evidence cases while increasing usable coverage by 20.8% through conservative augmentation.

Keywords: political elites; measurement; tool-using retrieval; replication; temporal leakage.

1 Introduction

Large-N political-elite datasets underpin research on representation, accountability, and governance, but assembling them requires repeated, error-prone measurement of categorical attributes (e.g., party affiliation) across countries and time. As digital traces proliferate, it is tempting to scale attribute coding by asking search-enabled large language models (LLMs) for answers. Yet, interactive question answering is not a measurement protocol: without explicit governance, such workflows tend to be difficult to audit, hard to replicate, and vulnerable to temporally invalid evidence.

The stakes are substantive, not only technical. Party affiliation and related elite attributes are widely used to define key covariates (e.g., partisan control, coalition membership, and ideological composition) and to subset samples (e.g., focusing on elected officials with identifiable party ties). When labels are missing or miscoded, downstream analyses can suffer selection bias and reduced comparability across countries and years; when labels are coded using post-period biographies or retrospective summaries, they risk post-t contamination of baseline covariates.

In political methodology terms, the problem is therefore *measurement design*: how do we scale retrieval-augmented coding of elite attributes while preserving (i) auditability (a third party can verify why a label was produced), (ii) cross-national comparability (labels remain in a stable, closed set of classes), and (iii) temporal validity (evidence respects an "as-of" target). ASA is framed as a measurement protocol rather than as an accuracy-maximizing question-answering system.

We argue that retrieval-augmented LLM agents can be useful for dataset construction only when embedded in an explicitly governed measurement protocol. We introduce the *Auditable Search Agent* (ASA), implemented in the `asa` software stack, which formalizes retrieval as measurement: for each record ⟨person, country, year⟩ and an expert-supplied closed codebook, the agent runs a short, budgeted search session; prioritizes evidence over parametric recall; emits a structured output with citations; abstains under weak or conflicting evidence; and archives a complete trace (queries, snippets, URLs, timestamps) for later audit and re-analysis.

Contributions. First, we identify three hazards that make "search-enabled coding" non-standard as a research workflow: (i) non-verifiability, (ii) non-comparability, and (iii) temporal leakage (a close analogue of post-treatment conditioning in causal inference). Second, we translate these hazards into a concrete protocol with closed-world decisions, evidence-first outputs, temporal governance, and conservative abstention. Third, we validate the methodology on party-affiliation labeling, a verifiable elite attribute with an expert closed codebook.

Preview of results. In the party-affiliation task spanning 114 countries (N = 34,618 matched leader–records; 1,209 party labels), ASA achieves high-confidence accuracy 0.860 while expanding usable coverage by 20.8% (12,898 previously missing labels) and withholding 0.250 of cases on average via an abstention gate. Relative to a simple country-majority baseline (0.536), the mean uplift in high-confidence accuracy is 0.343.

Roadmap. Section 2 motivates the design by formalizing three measurement hazards. Section 3 situates the contribution in political methodology, reproducible research, and recent retrieval-agent architectures. Sections 4–6 describe the ASA protocol and validate it on party affiliation. We conclude with guidance on transparency, limitations, and ethical considerations for web-based measurement. Our replication target is recomputation from a frozen trace store and versioned task specification; we do not claim that reruns on the live web will match.

2 Measurement hazards in search-enabled coding

Digital sources make it feasible to code political-elite attributes at scale, but research-grade datasets impose requirements that differ from interactive question answering. In practice, a default workflow—"ask a search-enabled model for the answer"—tends to fail on three dimensions central to quantitative political science.

Auditability and replication. Elite attribute labels are *measurements.* When the underlying retrieval context is not archived (queries, results, snippets, URLs, timestamps), a label is difficult to verify and cannot be re-audited when coders disagree or sources change. This is especially problematic for downstream users who need to understand measurement error and may wish to apply alternative inclusion rules (e.g., discarding any label lacking a primary source).

Cross-national comparability. Many elite variables are *categorical* and country-specific (e.g., party labels). "Open-world" generation invites label drift (synonyms, translations, rebrandings), undermining comparability across countries and years. Closed codebooks supplied by domain experts are a natural remedy, but generic search-enabled chat systems do not typically enforce them.

Temporal leakage and post-treatment bias. Political elites switch parties, offices, and coalitions. If coding for year t uses sources written after t (e.g., biographies updated in 2025), the measurement can incorporate information that is itself caused by post-t outcomes. This is a close analogue of conditioning on post-treatment variables in causal inference (Montgomery, Nyhan, and Torres, 2018): a label intended to reflect the pre-period state may be contaminated by later events, inducing bias in estimated relationships. Temporal governance therefore belongs *inside* the measurement protocol, not in ad hoc post-hoc cleaning.

2.1 Formalizing temporal leakage

Let X_{it} denote the target attribute for unit i at time t (e.g., party affiliation in observation year t). Researchers often use X_{it} as a covariate when estimating relationships with outcomes realized after t, $Y_{i,t'>t}$. Web-based coding observes a set of sources S_{it} whose content and availability depend on time: some sources are available at or before t ($S_{it}^{\leq t}$), while others are published or updated after t ($S_{it}^{>t}$). When a measurement procedure for X_{it} draws on $S_{it}^{>t}$, it can implicitly incorporate information downstream of $Y_{i,t'>t}$ (e.g., retrospective biographies updated after a party switch), creating look-ahead leakage and a close analogue of post-treatment conditioning.[1]

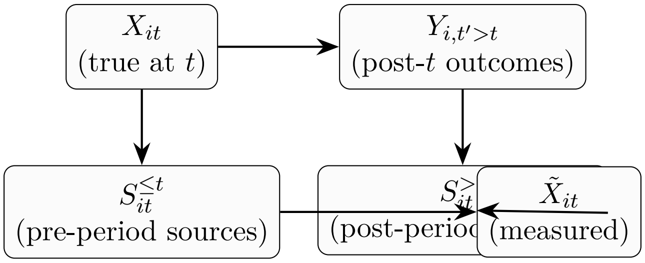

Figure 1: Temporal leakage as post-period conditioning. If measurement $\tilde{X}_{it}$ uses post-period sources $S_{it}^{>t}$ that are themselves influenced by outcomes after t, then $\tilde{X}_{it}$ can become a function of $Y_{i,t'>t}$, contaminating a baseline covariate.

[1] See Paleka et al. (2025) for a broader discussion of temporal validity pitfalls when "as-of" constraints are not enforced in evaluation settings.

A compliance definition. For each record i, the task specification includes a temporal cut τ_i (an "as-of" date). Each retrieved source $s \in S_{it}$ is assigned a best-effort publication (or last-update) date $d(s)$ when recoverable. Define the admissible evidence set $S_{it}^{\mathrm{adm}}(\tau_i) = \{s \in S_{it} : d(s) \text{ exists and } d(s) \leq \tau_i\}$. ASA treats a label as *temporally compliant* only if (i) its cited supporting sources are drawn from $S_{it}^{\mathrm{adm}}(\tau_i)$, and (ii) at least one cited supporting source has a recoverable $d(s)$. When no supporting source has a recoverable date, ASA downgrades evidence strength and abstains rather than silently relying on potentially post-period material.

Design implications. These hazards motivate treating retrieval as a governed measurement instrument: decisions must be constrained (closed codebooks), evidence must be preserved (citations and traces), time must be part of the protocol ("as-of" rules), and weak evidence should trigger abstention rather than guessing. Table 1 summarizes how ASA operationalizes these requirements.

Table 1: Common hazards in search-enabled coding and ASA design responses.

Hazard	**ASA response**
Non-verifiability (answers without durable evidence trails)	Evidence-first outputs with explicit citations; full trace capture (queries, snippets, URLs, timestamps) for audit and re-analysis.
Non-comparability (label drift across countries/time)	Closed-world decisions constrained to expert codebooks, plus conservative post-processing that normalizes benign surface variation without introducing new classes.
Temporal leakage (post-period information used for pre-period labels)	Explicit "as-of" rules with date verification where feasible; warnings and abstention rather than silently relying on potentially post-period sources.

2.2 Auditability and comparability as formal properties

Temporal leakage is only one way search-enabled coding can fail as measurement. To make retrieval an admissible instrument for political data construction, we also require (1) *auditability* (a third party can reconstruct why a label was emitted) and (2) *comparability* (outputs remain in a stable, closed set of classes across country-years).

The trace as the measurement object. Let Ω_t denote the (unobserved) state of the web at time t. A tool-using coding episode for record i produces a *trace*

$$T_i = \Big((a_{i1}, o_{i1}, \kappa_{i1}), \ldots, (a_{iK}, o_{iK}, \kappa_{iK})\Big),$$

where each step records the agent action a_{ik} (e.g., a query or page fetch), the observed tool output o_{ik} (e.g., ranked results and extracted snippets), and a timestamp κ_{ik}. The reported label is a function of the trace and the task specification:

$$\hat{X}_{it} = M(T_i; \mathcal{C}_i, \tau_i, E, \theta),$$

where $\mathcal{C}_i$ is the closed codebook, τ_i is the temporal cut, E is an evidence sufficiency rule, and θ denotes implementation parameters (e.g., mapping thresholds and source-tier rules).

Auditability. We say a coding procedure is *auditably reproducible* for a reported result if, given a *frozen* trace store $\mathcal{T} = \{T_i\}_{i=1}^{N}$ and a versioned task specification $(\mathcal{C}_i, \tau_i, E, \theta)$, an independent auditor can deterministically recompute every reported label and every aggregate statistic in the paper by replaying $M(\cdot)$ on $\mathcal{T}$. This is the relevant reproducibility target for web-based measurement: rerunning retrieval on the live web is not expected to match, but recomputation from an archived trace store is.[2]

Comparability via closed-world mapping. Cross-national comparability requires that categorical labels be stable and enumerable. ASA enforces a closed-world constraint by requiring $\hat{X}_{it} \in \mathcal{C}_i$ or abstention. Let $g(\cdot)$ map extracted surface strings from evidence into codebook entries. Closed-world comparability requires g to be *non-expansive*: it may normalize benign surface variation (punctuation, spacing, acronyms) but must never introduce a class outside $\mathcal{C}_i$. When g is undefined, or evidence supports multiple codebook entries, ASA abstains rather than generating a novel label. This operationalizes a basic measurement-validity principle: prefer explicit, auditable missingness to uncontrolled label drift that silently changes the meaning of a variable across units and time (**adcock2001measurement**; **munck2002conceptualizing**; **lazer2014parable**).

3 Related work

Our contribution sits at the intersection of political methodology and recent advances in retrieval-augmented language models. Large cross-national datasets often rely on expert coding and bespoke compilation efforts; a central challenge is preserving comparability and documenting measurement decisions at scale. ASA complements such efforts by treating web retrieval as a first-class part of the measurement protocol and by archiving evidence trails that allow downstream users to audit labels and re-apply inclusion rules. The validation task we study is motivated by the same large-N political-elite measurement agenda that underlies recent work on global leadership and expert-coded cross-national datasets (Gerring et al., 2019; Pemstein et al., 2020).

The political-science measurement literature emphasizes that validity hinges on transparent linkages between concepts, operational rules, and observed indicators, and that measurement choices must be documented in a way that supports reinterpretation and audit. ASA's closed codebooks, explicit abstention boundary, and trace-based audit target these concerns (**adcock2001measurement**; **munck2002conceptualizing**). More broadly, digital trace data can be alluring but unstable; classic warnings about shifting data-generating processes in "big data" settings apply directly to web-based measurement (**lazer2014parable**).

ASA is also aligned with transparency and replication norms in political science, where the goal is not merely to share code but to make measurement decisions inspectable and

[2] This distinction mirrors a broader point in agentic systems: raw action logs are often insufficient for claim-level auditability without structured, claim–evidence encodings (Rasheed et al., 2026).

contestable. Calls for data access and research transparency (DA-RT), "active citation," and practical guidance on code-and-data release motivate our emphasis on archivable evidence and replayable rules (**dart2015joint**; **moravcsik2014transparency**; **gentzkow2014code**).

Finally, ASA's abstention gate connects to a long literature on the error–reject tradeoff in recognition systems and to modern selective prediction work that treats abstention as a principled way to control risk (**chow1970reject**; **geifman2017selective**).

ASA is also aligned with calls for reproducible computational research: replication requires not only code, but durable documentation of the information environment that produced a label (Peng, 2011; Stodden, Guo, and Ma, 2013). In web-based measurement, the information environment is inherently dynamic; capturing queries, snippets, URLs, and timestamps makes it possible to re-audit outputs when sources change. Our emphasis on temporal governance connects directly to concerns about look-ahead and leakage when post-period information is used to measure pre-period covariates (Kaufman, Rosset, and Perlich, 2012; Montgomery, Nyhan, and Torres, 2018).

On the NLP side, retrieval-augmented generation and tool-using agents provide building blocks for ASA's implementation (Lewis et al., 2020; Yao et al., 2022). However, the core contribution here is not a new prompting trick or a larger model; it is a measurement-oriented design that prioritizes comparability, auditability, and temporal validity over unconstrained answer generation.

4 Methodology: an auditable search agent

We present a protocol—implemented as the `asa` software stack—that makes agent-based retrieval *verifiable by construction*. The methodology separates (i) a task-level specification that defines what constitutes admissible evidence and outputs, from (ii) an implementation that executes the specification and persists traces.

4.1 Protocol commitments

The protocol has four core commitments:

1. **Closed-world decisions.** All predictions must lie in an expert-supplied codebook (e.g., the parties observed in a country-year panel).
2. **Evidence-first outputs.** Each label is paired with a terse rationale anchored to explicit source citations.
3. **Provenance preservation.** The system archives the full interaction trace: queries, tool responses, extracted snippets, URLs, and timestamps.
4. **Conservative abstention.** Under weak or conflicting evidence, the agent abstains (or routes to humans) rather than guessing.

4.2 Abstention as selective measurement

ASA's abstention rule is not merely an implementation detail; it defines the statistical object that is being produced. Let Y_{it} denote the expert-coded party label (when available). ASA returns a *selective* prediction $\hat{Y}_{it} \in \mathcal{C}_i \cup \{\perp\}$, where $\perp$ denotes abstention. Two quantities therefore matter jointly:

$$\text{Coverage } \gamma \equiv \Pr(\hat{Y}_{it} \neq \perp), \qquad \text{Conditional error } R \equiv \Pr(\hat{Y}_{it} \neq Y_{it} \mid \hat{Y}_{it} \neq \perp).$$

Tightening evidence requirements mechanically reduces R while lowering γ, an instance of the classic error–reject tradeoff (**chow1970reject**) and closely related to modern selective classification (**geifman2017selective**). In elite measurement settings, abstention is attractive because withheld cases can be routed to human coders or left missing under transparent, replayable rules—a preferable failure mode to un-auditable guesswork.

4.3 Task formalization

For each target record i, the inputs are: a person name, a target country, an observation year t_i, and a closed codebook $\mathcal{C}_i$ supplied by expert coders. The agent must return either (a) a label $c \in \mathcal{C}_i$ with citations, or (b) an abstention.

Protocol 1 (ASA): governed retrieval as measurement.

1. **Inputs.** Record $r_i = \langle \text{person, country, year} t_i \rangle$; expert codebook $\mathcal{C}_i$; temporal cut τ_i ("as-of"); retrieval budget B; and an evidence sufficiency rule E (task-specified).
2. **Retrieve (trace everything).** Issue up to B read-only retrieval actions (e.g., Wikipedia + web search). Persist each query, ranked results, extracted snippets, URLs, and timestamps.
3. **Date and filter sources.** For each candidate source s, recover a publication/last-update date $d(s)$ when feasible. Mark sources with $d(s) > \tau_i$ as post-period and inadmissible for a high-evidence decision; mark sources without recoverable $d(s)$ as undated.
4. **Extract candidate labels.** From admissible snippets, extract party-name strings and map to $\mathcal{C}_i$ via strict match, then conservative fuzzy normalization for benign surface variation.
5. **Decide or abstain.** If mapped evidence meets E using admissible sources and does not conflict, output $c \in \mathcal{C}_i$ with a one-sentence, citation-backed justification and an evidence-strength tier. If evidence is weak, conflicting, or only undated, abstain.
6. **Persist artifacts.** Write the structured output (including abstentions) and a pointer to the full trace store entry so results can be re-audited and aggregate statistics recomputed from the frozen trace store.

Figure 2: A submission-ready ASA protocol statement. The task configuration specifies $\mathcal{C}_i$, τ_i, B, and E; the shared execution layer enforces trace capture, temporal admissibility, and abstention.

4.4 Evidence sufficiency and confidence tiers

The protocol parameter E encodes what counts as *sufficient evidence* for emitting a high-confidence label. In party-affiliation coding, we implement E as a tiered evidence rule that privileges primary, contemporaneous sources and requires independence across sources.

Source tiers. Each retrieved source s is assigned a coarse reliability tier—primary, secondary, or tertiary—using domain lists and page cues. We treat *primary/official* sources (e.g., government, parliament, election commission, party roster pages) as the strongest evidence; *secondary* sources include major news outlets and established reference works; *tertiary* sources (e.g., scraped bios and low-editability directories) may be used for navigation but are not, by themselves, sufficient for high-evidence decisions.

Independence. To reduce correlated error, we treat two sources as independent if they come from different base domains and are not obvious mirrors or syndicated copies. This criterion is deliberately coarse but easy to audit in stored traces.

Tiered sufficiency rule. Let $S_{it}^{\mathrm{adm}}(\tau_i)$ be the admissible (dated, pre-τ_i) evidence set defined in Section 2.1. ASA assigns each record to one of three evidence tiers:

1. **Tier A (high evidence):** At least two independent admissible sources map non-expansively to the same codebook label, including at least one primary/official source with a recoverable date.
2. **Tier B (moderate evidence):** Either (i) one admissible primary/official source maps cleanly to a label, or (ii) two independent secondary sources agree, but primary evidence is unavailable.
3. **Tier C (weak/undated/conflicting):** Evidence is undated, post-period, maps to multiple labels, or relies only on tertiary sources.

In our default configuration we emit a high-confidence label only under Tier A; Tier B is available for human review or sensitivity checks; Tier C cases are abstentions. This makes the abstention boundary explicit and replayable: downstream users can tighten or relax E and recompute headline statistics from the same frozen trace store.

Table 2: Evidence tiers used by the sufficiency rule E in party-affiliation coding.

Tier	**Minimum sources**	**Operational rule (summary)**
A (high evidence)	≥ 2 independent	Dated, admissible sources agree on the mapped codebook label; includes ≥ 1 primary/official source with recoverable date; no unresolved conflict. Emit a high-confidence label.
B (moderate evidence)	1 primary or 2 secondary	Agreement under admissible dating, but primary evidence is limited or unavailable. Route to human review or treat as medium confidence in sensitivity checks.
C (weak / conflicting)	–	Evidence is undated or post-period, relies on tertiary sources, or supports multiple labels with comparable strength. Abstain.

4.5 Conflict policy: resolve or abstain

Conflicts arise when two different codebook labels are each supported by at least one independent admissible source. Because post-hoc tie-breaking can silently encode researcher discretion, ASA defaults to abstention under conflict unless one label strictly dominates by evidence tier (e.g., Tier A support versus Tier B/C) and source tier (primary versus secondary/tertiary). All conflict flags and supporting citations are persisted so disagreements are auditable and can be routed to human coders.

4.6 Entity disambiguation

Elite names are not unique: homonyms, transliterations, and office changes can induce systematic misattribution if not governed explicitly. ASA therefore requires that at least one cited evidence span contain a corroborating identifier beyond the name (e.g., office/chamber, country, year/tenure, or constituency). When disambiguation remains unresolved, ASA abstains rather than guessing.

4.7 Implementation: ASA software stack

Operationally, the agent runs a short search session with read-only retrieval tools (e.g., Wikipedia and general web search), compiles candidate evidence, and emits a structured JSON result (label, one-sentence justification with citations, and a confidence category). This design follows retrieval-augmented, tool-using agent patterns in the recent NLP literature (Lewis et al., 2020; Yao et al., 2022). The implementation is designed to be *task-configurable*: domain experts supply the closed codebook and task schema, while the shared execution layer enforces the protocol commitments (trace capture, abstention, and evidence-first outputs).

A conservative post-processor then normalizes the label against the codebook: exact-match acceptance first, followed by a guarded fuzzy match for benign surface variation (punctuation, plurals, acronyms) using high similarity thresholds. This reduces typographic drift without introducing new classes; in the party-affiliation case study, the relaxed mapper changes 2.5% of accepted labels.

4.8 Temporal governance to prevent leakage

The agent treats time as part of the measurement protocol. Each run is parameterized by an "as-of" rule: sources should be published before a cut date (e.g., $t_i + 1$ year) or within a target window around t_i. When feasible, the system extracts publication dates from retrieved pages and enforces the constraint strictly; otherwise it falls back to best-effort warnings and abstention rather than silently using potentially post-period material. This guards against a common dataset-construction failure mode: coding pre-period attributes with post-period knowledge.

Operational date recovery $d(s)$. For each retrieved page, ASA attempts to recover a publication or last-update date using an ordered set of heuristics: (i) structured metadata (e.g., schema.org JSON-LD fields such as `datePublished/dateModified`), (ii) HTML meta

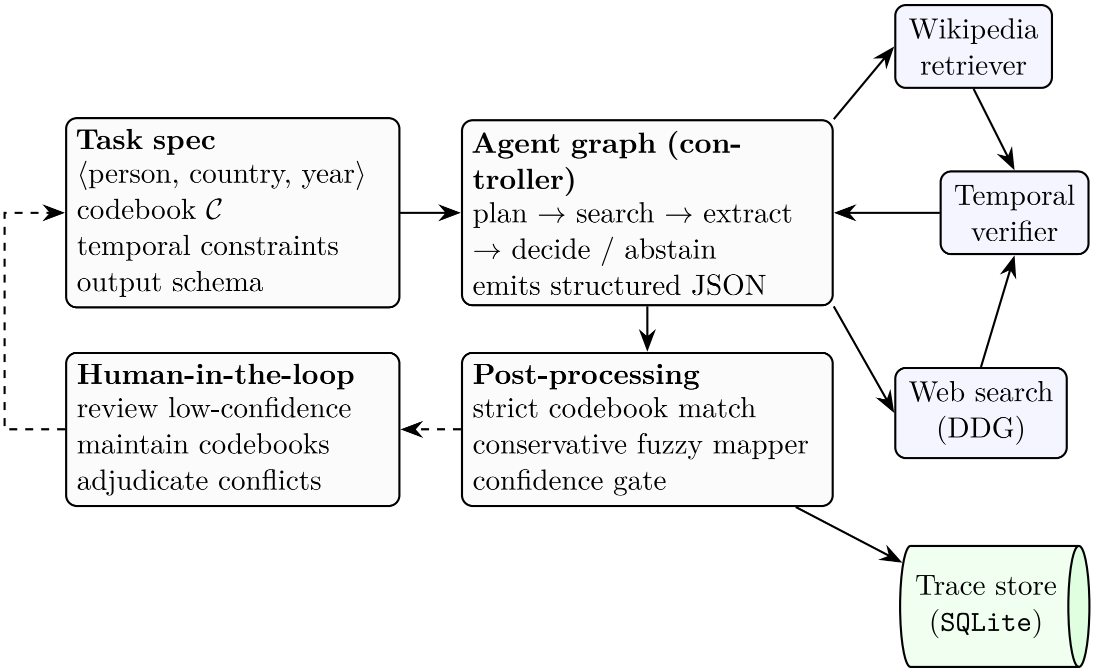

Figure 3: Design visualization: a governed agent graph with temporal verification, conservative normalization, abstention, and an auditable trace store.

tags (e.g., OpenGraph/RSS cues), (iii) visible byline dates in page text, and (iv) weak URL date patterns as a last resort. If no reliable date is found, the source is treated as undated and cannot be the sole support for a Tier A decision (Section 4.4).

Temporal controls are crucial because many online sources are retrospective and continuously updated: Wikipedia leads, later biographies, and obituaries often summarize whole careers, including events after t_i. If such post-t_i material is used to code a pre-period attribute, the measurement can incorporate information that is itself downstream of later outcomes (e.g., a party switch or coalition realignment that occurs after t_i), creating look-ahead/data leakage and a close analogue of post-treatment conditioning. This can both inflate apparent coding accuracy (by using information unavailable at the time) and bias downstream relationships by contaminating "baseline" covariates with post-period events (Kaufman, Rosset, and Perlich, 2012; Montgomery, Nyhan, and Torres, 2018).

4.9 Provenance preservation and storage

All tool interactions and intermediate artifacts are persisted (including URLs and timestamps) to enable: (i) verification of any individual label by following its citations, (ii) replication of aggregate statistics given the same trace store, and (iii) retrospective audits when codebooks change or new sources appear. The storage design also supports alternative downstream rules (e.g., only keep labels supported by government sources).

4.10 Trace schema and audit procedure

A key claim of ASA is that it produces measurements that are *auditable by construction.* To make this concrete, the trace store persists two linked objects per record i:

(i) A structured decision record. This is the minimal unit needed to reproduce the paper's tables:

$$D_i = \{\text{record id},\ \hat{Y}_{it},\ \text{tier},\ \mathcal{C}_i,\ \tau_i,\ \text{supporting citations},\ \text{conflict flags},\ \text{software version}\}.$$

The "supporting citations" field points to specific snippet spans within retrieved pages (not only URLs), so an auditor can verify that the cited text supports the mapped codebook label.

(ii) A full interaction trace. For each tool call k, we store: query string or URL, ranked results, bounded snippet text, timestamp, domain, recovered publication/update date (if any), and extraction metadata. Storing bounded snippets rather than full pages reduces privacy and copyright exposure while preserving enough context for audit. Appendix **??** summarizes the minimum field set.

Audit checklist (replayable). Given (D_i, T_i), an auditor can verify a high-confidence label by checking: (1) all cited sources are admissible under the temporal cut ($d(s) \leq \tau_i$), (2) cited spans contain an explicit party mention that maps non-expansively into $\mathcal{C}_i$, (3) the required number of independent sources is met for the claimed tier (Section 4.4), and (4) there is no unresolved conflict with similarly supported alternatives (Section 4.5). This moves verification from "trust the model" to "inspect the stored evidence."

5 Data and gold standard

We validate ASA on party affiliation, a verifiable elite attribute that is frequently missing in large political-elite panels yet can often be corroborated from public records (party rosters, parliamentary biographies), reputable encyclopedic sources, and contemporaneous news coverage. The unit of analysis is a leader–record indexed by ⟨person, country, year⟩.

5.1 Leader records, temporal scope, and unit construction

We begin from a cross-national panel of political elites assembled as part of the Global Leadership Project research program (**gerring2024composition**; Gerring et al., 2019). From this source we construct leader–records indexed by ⟨person, country, year t⟩. The observation year t defines the temporal target for party affiliation, and each ASA run is parameterized by a record-specific temporal cut τ_i (Section 2.1).

Because names vary across languages and sources, we retain both the original name string and a normalized form (case-folding, diacritic handling) in the trace store. This supports downstream audits of homonyms and transliterations (Section 4.6).

5.2 Expert labels and codebook construction

For each country-year, domain experts provide a *closed codebook* $\mathcal{C}_i$ of admissible party labels. The codebook enforces cross-national comparability by preventing open-ended label generation, while allowing country-specific party systems to be represented without forcing artificial cross-country harmonization. Operationally, codebooks are versioned and frozen for an ASA run; when party naming conventions evolve (coalitions, mergers, rebrands), codebooks can be updated and prior labels can be re-audited using archived traces. To reduce conservative normalization error, codebooks may include documented aliases that map non-expansively into canonical labels (Section 2.2).

Expert party labels used for evaluation come from the underlying expert-coding workflow. We treat these labels as the gold standard for validation, while emphasizing that ASA's trace store preserves enough evidence to support audits and disagreement adjudication, consistent with measurement best practices (**adcock2001measurement**; **munck2002conceptualizing**).

5.3 Record linkage and match quality

Leader records, expert labels, and country-year codebooks must be linked to construct the evaluation subset. Where unique identifiers are available, linkage is deterministic; otherwise we implement a conservative matching procedure based on normalized names plus country and year, routing ambiguous matches to manual review. This conservative linkage step is important: linkage error can otherwise be misattributed to the coding agent. We preserve linkage decisions and match flags in the trace store so they are auditable (**fellegi1969theory**).

5.4 Outcome definition and temporal target

The target label is the leader's party affiliation for observation year t_i. Because elites may switch parties within a year and sources are often retrospective, ASA applies an explicit "as-of" rule for admissible evidence. The precise temporal cut and adjudication rules for within-year switches should be stated alongside the task configuration (Appendix **??**).

6 Validation: party affiliation labeling

6.1 Experimental design

Each record triggers a short, budgeted retrieval episode. The agent is instructed to prioritize verifiable sources, provide citations in its one-sentence justification, and abstain (low confidence) when evidence is weak or conflicting. Temporal governance is applied as described in Section 4: where publication dates can be recovered, post-period sources are excluded; otherwise the agent warns and abstains rather than silently relying on potentially post-period material.

6.2 Baselines and ablations

We report performance against a simple baseline that predicts, within each country, the modal party label in the expert-coded evaluation subset. This captures how much performance comes from exploiting skewed label distributions. In addition, a standard submission-ready validation should include: (i) an open-world "search-enabled chat" baseline without closed codebooks, (ii) a no-temporal-governance ablation, (iii) a no-abstention ablation (forced prediction), and (iv) a strict-match-only ablation (no relaxed mapper). We provide protocol details for these comparisons in Appendix **??**.

6.3 Metrics

Because ASA is a selective measurement instrument (Section 4.2), we report both performance *conditional on non-abstention* and the corresponding coverage. Let $\gamma \equiv \Pr(\hat{Y}_{it} \neq \perp)$ denote coverage (the share of records receiving a non-abstained label), and let $R \equiv \Pr(\hat{Y}_{it} \neq Y_{it} \mid \hat{Y}_{it} \neq \perp)$ denote conditional error. Our headline "high-confidence accuracy" is $1 - R$ computed on the Tier A/high-evidence subset under the default evidence rule E (Section 4.4). We also report overall accuracy across all emitted labels and coverage gain relative to expert-only availability. Because the unit of analysis is nested within countries, uncertainty should be summarized with country-clustered or country-resampled intervals in a submission-ready version of the paper.

6.4 Threats to validity in evaluating web-based measurement

Two evaluation design choices are essential to interpret ASA's headline numbers.

First, accuracy is reported *conditional on non-abstention*. This is appropriate for a selective measurement instrument, but it means that evaluation is performed on the subset of records that satisfy the evidence sufficiency rule E under the temporal cut. As a result, accuracy alone is not meaningful without the corresponding coverage γ and a description of which cases are withheld (e.g., early years with sparse web records, low-salience elites, or non-English contexts).

Second, web retrieval creates a potential dependence between the information environment available to experts and to ASA. To avoid conflating "shared sources" with methodological success, ASA stores its full evidence trails: readers can directly inspect whether correct labels are supported by contemporaneous primary sources or by retrospective summaries. This is one reason temporal governance is a core part of the protocol rather than a post-hoc evaluation filter.

Finally, several error modes are predictable in elite party coding: homonyms and transliterations (addressed via disambiguation, Section 4.6); within-year party switches (addressed via record-specific temporal cuts, Section 5.4); coalition or alias naming (addressed via non-expansive codebook mapping, Section 2.2); and sources without reliable dates (addressed via date recovery and abstention, Section 2.1).

6.5 Results

Using the subset of leader–records with expert codings, ASA attains an overall high-confidence accuracy of 0.860 across 114 countries ($N = 34{,}618$), while expanding coverage by 20.8% (12,898 records) through conservative augmentation. Low-confidence outputs are withheld by design, trading recall for precision; the mean share withheld is 0.250. Table **??** summarizes the key quantities.

Table 3: Validation summary for party-affiliation labeling (high-confidence predictions scored against expert labels).

Quantity	**Value**
Countries	114
Matched leader–records (expert overlap)	34,618
Distinct party labels (closed codebooks)	1,209
High-confidence accuracy (non-abstained)	0.860
Overall accuracy (all predictions)	0.751
Mean share withheld (low confidence)	0.250
Mean country-majority baseline accuracy	0.536
Mean uplift over baseline (high confidence)	0.343
Coverage gained (new labels)	20.8% (12,898 records)
Accuracy for small/minority parties	0.692
Accuracy for large/plurality parties	0.820

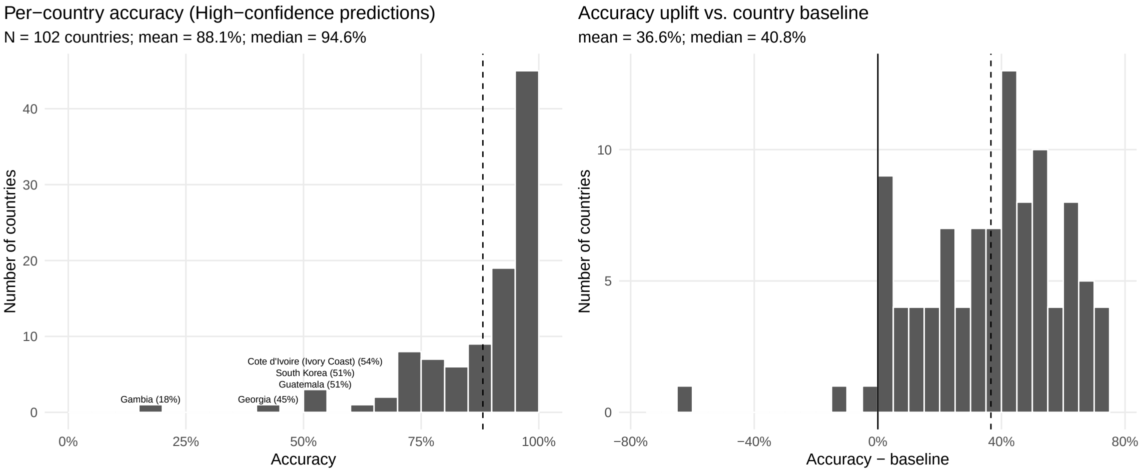

Figure 4: Agent performance in predicting party across countries with sufficient expert overlap. High-confidence predictions are evaluated against expert labels.

6.6 Auditability and trace-based verification

The core advantage of ASA over unconstrained chat workflows is that each accepted label can be audited by inspecting its citations and full trace. The trace store preserves the queries,

Table 4: Sample agent traces. Text content truncated for readability (and may contain typograical errors as present in native source). Links are clickable. Full traces contain many more sources.

Field	**Content**
Entry 1: Syleiman Abusaidovich Kerimov — Russian Federation (1999)	
Country	Russian Federation
Year	1999
Person	Syleiman Abusaidovich Kerimov
Wikipedia	Page: Ashot Egiazaryan Summary: Ashot Gevorkovich Egiazaryan (Russian: Ashot Gevorkovich Egiazaryan; Armenian: Ashot Gevorgovich Ekyazaryan; born...
Search 1	In the spring of 1998, Yeltsin dismissed Chernomyrdin as head of government and in1999Yeltsin's administration backed a newly formedparty,Un...
URL 1	https://en.wikipedia.org/...
Search 2	OURHOMEISRUSSIAPARTYOurHomeIsRussia(Nash Dom—Rossiya, or NDR) was a sociopolitical movement and a rulingpartyfrom 1996 to 1998. Source for i...
URL 2	https://www.encyclopedia.com/...
Entry 2: Jasminka Stanojevic — Serbia (2018)	
Country	Serbia
Year	2018
Person	Jasminka Stanojevic
Wikipedia	Page: Supreme Court (Serbia) Summary: The Supreme Court (Serbian: Vrhovni sud, romanized: Vrhovni sud) is the court of last resort in Serbia...
Search 1	This article lists political parties inSerbia, including parties that existed in the Kingdom ofSerbiabetween the early 1860s and 1918. A kol...
URL 1	https://en.wikipedia.org/...
Search 2	Imali su dve i četiri godine kad smo izbegli iz Knina. Kad bi neko pokucao na vrata, vikali bi: „Tata, tata". Tri godine nakon progona sazna...
URL 2	https://www.kurir.rs/...
Entry 3: Mihai STROE — Romania (2011)	
Country	Romania
Year	2011
Person	Mihai STROE
Wikipedia	Page: Adrian Stroe Summary: Adrian Stroe (born 24 October 1959), known as The Taxi Driver of Death, is a Romanian serial killer responsible ...
Search 1	Născut în Bucureşti şi cu origini în comuna argeşeană Morăreşti, fost medaliat cu aur la olimpiada internaţională de informatică,MihaiStroe(...
URL 1	https://adevarul.ro/...
Search 2	MihaiSTROEParliamentary activity in legislature 2008-2012 DEPUTY Constituency no.38 TULCEA, uninominal college no.2 Membru al PDL, deputatul...
URL 2	https://www.cdep.ro/...
Entry 4: Matsie Angelina Motshekga — South Africa (2018)	
Country	South Africa
Year	2018
Person	Matsie Angelina Motshekga
Wikipedia	Page: Angie Motshekga Summary: Matsie Angelina "Angie" Motshekga (born 19 June 1955) is a South African politician and educator who is curre...
Search 1	MatsieAngelina"Angie"Motshekga(born 19 June 1955) is a SouthAfricanpolitician and educator who is currently serving as the Minister of Defen...
URL 1	https://en.wikipedia.org/wiki/Angie_Motshekga
Search 2	Motshekgawas elected thenationalpresident of theAfricanNationalCongressWomen's League (ANCWL) in 2008, defeating the League's secretary-gene...
URL 2	https://www.sahistory.org.za/people/matsie-angelina-motshekga-angie-motshekga

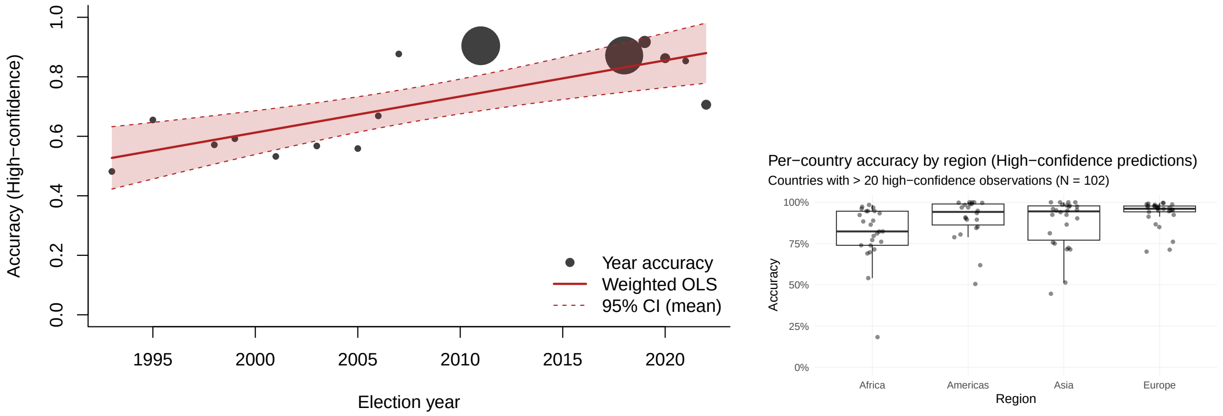

Figure 5: Performance heterogeneity by time (left) and region (right).

retrieved snippets, URLs, and timestamps used during coding, enabling replication and retrospective re-analysis when sources change or when users adopt stricter inclusion rules (e.g., accept only government sources). Table **??** illustrates the resulting evidence-first trace style: multiple sources are retrieved and recorded alongside links, allowing a reader to verify what evidence the label relied on under the declared temporal rule.

Worked example: why "as-of" governance matters. Consider a record with observation year t_i and a label intended to reflect the pre-period state (e.g., party affiliation at t_i). A naive workflow that reads a current biography or Wikipedia lead may inadvertently use retrospective summaries that incorporate events after t_i, creating look-ahead/data leakage (Kaufman, Rosset, and Perlich, 2012) and a close analogue of post-treatment conditioning (Montgomery, Nyhan, and Torres, 2018). ASA instead treats time as part of the measurement protocol: it constrains retrieval to sources published before a cut date (or within a target window), and it abstains when the available evidence cannot be shown to be pre-period. Table **??** illustrates the resulting trace style: multiple sources are retrieved and recorded alongside URLs, allowing a reader to verify what evidence the label relied on under the declared temporal rule.

Figure 6: A compact intuition for temporal governance and trace-based verification.

7 Downstream payoff: reducing selection and stabilizing inference

ASA is designed to improve measurement quality, but its practical value for political science is ultimately downstream: missing or temporally invalid elite attributes induce selection and measurement error in the analyses that use them. Party affiliation is a basic input to common derived quantities (e.g., party fractionalization within a legislature-year) and is routinely used as a covariate or sample filter. When party labels are sparse, analysts either (i) restrict to a smaller expert-coded subset, which can change the country-year composition of the sample, or (ii) accept larger but weakly documented labels that are difficult to audit. ASA's conservative augmentation is intended to expand usable coverage while keeping an explicit abstention boundary.

To illustrate the inferential stakes, we re-estimate a representative cross-national specification in which party fractionalization is a key covariate and governance outcomes are the dependent variables (Appendix ??). Table ?? summarizes the resulting change in sample composition and a headline coefficient when moving from an expert-only construction to expert plus ASA augmentation (using the same coverage threshold and abstention policy).

Table 5: Illustrative downstream impact of ASA augmentation. Summary drawn from the democracy specification in Appendix ??.

Party-label source	**Countries**	**Observations**	**Party frac. coefficient (t)**
Expert only	110	162	0.38 (3.81)
Expert + ASA (abstain)	135	224	0.55 (6.97)

The key point is not the substantive interpretation of any single coefficient, but that measurement choices about party labels change which countries and years enter standard empirical models. ASA makes this tradeoff explicit and replayable: the augmented sample is larger, and every additional label is accompanied by citations and a stored trace that can be audited or re-filtered under stricter evidence rules.

8 Discussion: design tradeoffs and practical guidance

Cost and scaling. The methodology is built for high throughput: each record triggers a short, budgeted tool-usage episode and yields a structured output that can be scored, filtered, and aggregated without manual parsing. Using smaller instruction-tuned models for orchestration, caching retrieval outputs, and routing only uncertain cases to humans yields substantial cost savings relative to fully manual coding or repeated interactive browsing.

What this approach does (and does not) guarantee. Provenance capture makes the *evidence* for each label inspectable, but it does not magically eliminate ambiguity in the world. The point of abstention is to keep ambiguity from silently becoming noise. Similarly, temporal governance reduces leakage risk, but cannot fully recover historical truth when the web record is sparse. In practice, auditability means a downstream user can inspect a label by following its recorded citations and can re-apply alternative inclusion rules (e.g., require government sources only) using the archived trace, consistent with calls for reproducible computational work (Peng, 2011; Stodden, Guo, and Ma, 2013). Key failure modes include (i) sources without reliable publication dates, (ii) sparse historical web records for earlier years, and (iii) entities with name ambiguity across languages; ASA responds by warning, tightening evidence requirements, and abstaining rather than imputing.

When to use ASA. ASA is best suited to *verifiable attributes* with authoritative sources (party membership, office holding, election outcomes) and stable codebooks. For sensitive or non-verifiable attributes (e.g., ethnicity without self-identification), we recommend stricter abstention and explicit ethical review.

9 Transparency, reproducibility, and data availability

ASA produces two primary research artifacts: (i) a structured dataset of labels (including abstentions and confidence flags), and (ii) an auditable trace store that documents the evidence environment that produced each label. To support replication, we distinguish two targets: (i) recomputing reported metrics from a *frozen trace store* (fully reproducible), versus (ii) rerunning fresh retrieval on the contemporary web (not expected to match). Concretely, a submission-ready package should (a) snapshot the trace store used for reported results, (b) include scripts that reproduce all tables and figures from that snapshot, and (c) version the task specification (codebooks, temporal rules, output schema) that governed retrieval. Where legally feasible, key sources can be augmented with lightweight stability aids (e.g., content hashes or archived snapshots) so that trace-based audits remain possible even under link rot (**klein2014scholarly**; **rfc7089**). More generally, recent work emphasizes that raw logs are necessary but not sufficient for claim-level auditability in agentic systems, motivating structured traces that make evidence explicit and machine-checkable (Rasheed et al., 2026).

Replication package contents (JOP). In line with Journal of Politics reproducibility expectations (**jop_instructions**), the replication package for this article will include: (i) a `README` specifying software versions and execution order, (ii) the analysis dataset(s) used to generate every table and figure in the main text and appendix, (iii) a codebook describing every variable used, and (iv) scripts that reproduce the paper from a frozen trace store snapshot (i.e., recompute all accuracy/coverage statistics and regenerate all figures without re-querying the live web). Where full trace release is constrained (e.g., terms of service, privacy, or copyright), we will provide de-identified example traces, the complete schema, and an access pathway for qualified auditors to verify claims against the full trace store under controlled conditions.

Because the underlying elite panel and full trace store may be subject to access constraints, we recommend a controlled-access transparency strategy: release the `asa` software, schema definitions, and de-identified example traces immediately; and provide reviewers (and later readers) with an access pathway for full traces and expert codebooks consistent with ethical and legal constraints.

10 Limitations and ethical/legal considerations

Dynamic sources and dating. Web sources evolve (link rot, edits, retroactive updates), and many pages lack reliable publication dates. Temporal governance reduces leakage risk but cannot guarantee that all evidence is contemporaneous; abstention is therefore a feature, not a bug. These concerns are not hypothetical: reference rot is widespread in scholarly communication, motivating both trace preservation and, where feasible, time-based access/snapshotting approaches (**klein2014scholarly**; **rfc7089**).

Coverage and representation bias. Evidence quality varies systematically by language, region, and historical period, and across types of elites. These differences can create uneven missingness and error that must be documented and, where possible, modeled downstream. Because ASA abstains under weak evidence, non-abstained labels are not a random subset of records; coverage γ therefore functions as a selection mechanism that should be reported and stress-tested by country, time, and elite type.

Terms of service and privacy. Web retrieval should respect site terms, robots policies, and privacy considerations. Trace storage policies should minimize unnecessary retention of personal data while preserving enough evidence for audit (e.g., store URLs and short snippets rather than bulk page content when feasible).

Scope and sensitive attributes. ASA is best suited to verifiable attributes with authoritative sources (party membership, office holding, election outcomes) and stable codebooks. For sensitive or non-verifiable attributes (e.g., ethnicity without self-identification), we recommend stricter abstention, explicit ethical review, and transparency about uncertainty and potential harm.

11 Conclusion

For political-elite research, the central question is not whether LLMs can answer factual questions, but whether they can be integrated into a *measurement protocol* that is auditable, comparable, temporally well-defined, and cost-effective at scale. The Auditable Search Agent protocol operationalizes these requirements through closed codebooks, evidence-first outputs, strict trace preservation, and conservative abstention. This design turns retrieval-augmented agents into replicable instruments for dataset construction rather than opaque assistants.

Competing interests

The authors declare no competing conflict of interest.

Author contributions

TBD.

A Task specification for party affiliation

This appendix documents the task-level specification used in the party-affiliation validation: the required output schema, the country-closed normalization rules, the confidence gate (abstention policy), and a representative prompt template.

A.1 Structured output schema

For each record ⟨person, country, year⟩ the agent returns a single JSON object. In the party-affiliation task, we require (at minimum) the following fields:

- `pol_party`: a single party label chosen from the country-year codebook (exact string match).
- `justification`: one sentence that cites the evidence supporting the label.
- `confidence`: a categorical confidence flag (High, Medium, Low).

The trace store separately records the evidence environment (queries, snippets, URLs, timestamps) used to generate the output.

A.2 Country-closed matching and normalization

To guard cross-national comparability and reduce typographic drift, we apply a two-stage, codebook-guided normalization to the model's raw label string:

1. **Strict closed-set match (country scope).** If the predicted string exactly matches a member of the country-year codebook, it is accepted.
2. **Conservative fuzzy match.** Otherwise, a relaxed mapper computes a similarity score s combining (a) Jaro–Winkler similarity on a normalized label, (b) token overlap coverage, and (c) acronym equality. Let m denote the runner-up score. Accept as a match if

$$s \geq 0.92 \quad \text{or} \quad \Big(s \geq 0.85 \;\&\; s - m \geq 0.08\Big).$$

This mapping process produces a normalized label used for evaluation and aggregation while preserving the original string for audit. It corrects innocuous variants (pluralization, punctuation, acronyms) without introducing new classes. In the party-affiliation case study, the relaxed mapper changes 2.5% of accepted labels.

A.3 Confidence gate and abstention policy

The agent emits a categorical confidence estimate. Records flagged `Low` (and, in conservative downstream analyses, `Medium`) are withheld from automated use, implementing an abstention layer that trades coverage for precision. Abstention is most common where web evidence is sparse, parties are newly formed, transliterations vary, or publication dates cannot be verified under the "as-of" rule.

A.4 Representative prompt template

```
TASK OVERVIEW:
You are a search-enabled language model performing
party affiliation inference.
Your goal is to identify the political party of a
specified individual in a specified country and year,
using retrieved evidence. If evidence is weak or
conflicting, you must abstain by returning
confidence = "Low".

TARGET RECORD:
- Name: <PERSON_NAME>
- Country: <COUNTRY>
- Observation year: <YEAR>
- Parties (closed codebook): [<PARTY_1>, <PARTY_2>, ...]

CONSTRAINTS:
1. You MUST choose exactly ONE party from the closed
   codebook for pol_party when confidence is High or
   Medium.
2. You MUST NOT invent a party not in the codebook.
3. The selected party string must exactly match the codebook entry.
4. Write all explanations in English and cite sources.

RESPONSE FORMAT (JSON ONLY):
{
  "justification": "One sentence with citations to retrieved sources.",
  "pol_party": "Exact party label from the codebook",
  "confidence": "High|Medium|Low"
}
```

B Recommended additional validation for submission

To meet common journal expectations for a submission-ready measurement paper, the validation section should include:

- **Open-world baseline:** a “search-enabled chat” workflow without closed codebooks and without trace enforcement.
- **No temporal governance ablation:** remove the “as-of” constraint and show the effect on accuracy/coverage (and leakage risk).
- **No abstention ablation:** force a label on all records and report the precision–coverage tradeoff.
- **Strict-match-only ablation:** remove the relaxed mapper to quantify the role of normalization.
- **Uncertainty summaries:** country-clustered or country-resampled intervals for headline metrics.

C Trace store fields

The ASA trace store is designed to make each measurement auditable. At minimum, it records (a) the record identifier and task configuration (including the codebook and temporal rule), (b) each retrieval action (query strings, tool responses, extracted snippets), and (c) the final structured output with confidence. This enables retrospective audits (follow the citations), re-filtering (apply stricter evidence rules), and reproducible aggregate statistics from a frozen trace store.

Table 6: Minimum trace store fields needed for replayable audit and recomputation.

Object	**Required fields (minimum)**
Decision record D_i	Record identifier; emitted label or abstention; evidence tier; codebook identifier/version; temporal cut τ_i; supporting citations (URLs plus bounded excerpt text); conflict flags; software version; run timestamp.
Interaction trace T_i	Tool-call id; tool type; query string or URL; ranked results (bounded); extracted snippet/excerpt text (bounded); source domain; recovered publication/update date (and confidence) when available; tool timestamps; extraction metadata (e.g., parse method).

D Downstream illustration: full results

This appendix provides the full regression output underlying the downstream illustration in Section **??**. The key comparison is between models estimated on (i) an expert-only party-label construction and (ii) an expert plus ASA augmentation that retains an explicit abstention boundary. The primary purpose is to show that party-label measurement choices change sample composition (countries and legislature-years) and can therefore change applied inferences, even when headline coefficients point in similar directions.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Outcome	Democracy	Democracy	Corruption	Corruption	Pol. Stability	Pol. Stability
Segmentation Index		-0.26 (-1.45)		-0.72 (-0.61)		-1.09 (-1.76)
Group Frac.	0.01 (0.20)	0.12 (1.19)	-0.05 (-0.10)	0.26 (0.33)	-0.02 (-0.08)	0.44 (1.29)
Party Frac.	0.38 (3.81)*	0.45 (3.83)*	0.37 (0.76)	0.55 (1.08)	-0.03 (-0.15)	0.25 (1.10)
log(GDP p.c., PPP)	0.09 (3.91)*	0.09 (3.83)*	0.85 (6.76)*	0.85 (6.81)*	0.50 (6.96)*	0.50 (7.01)*
log(Population)	-0.03 (-2.25)*	-0.03 (-2.20)*	-0.14 (-2.24)*	-0.14 (-2.21)*	-0.20 (-5.62)*	-0.20 (-5.40)*
Percent Urban	0.00 (1.21)	0.00 (1.09)	0.01 (0.85)	0.01 (0.82)	0.00 (0.07)	0.00 (-0.05)
Other statistics						
Countries	110	110	110	110	110	110
Observations	162	162	162	162	164	164
Adjusted R-squared	0.46	0.47	0.53	0.53	0.61	0.62

Table 7: Elite-level ethnic cleavage and governance outcome. OLS with stand errors clustered at the group level. $*$ indicates $p < 0.05$; t-statistics in parentheses.

References

Gerring, John et al. (2019): "Who rules the world? A portrait of the global leadership class". In: *Perspectives on politics*, no. 4, vol. 17, pp. 1079–1097.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Outcome	Democracy	Democracy	Corruption	Corruption	Pol. Stability	Pol. Stability
Segmentation Index		-0.29 (-1.42)		-1.28 (-1.14)		-2.43 (-3.62)*
Group Frac.	0.03 (0.43)	0.13 (1.31)	0.12 (0.33)	0.58 (0.98)	0.28 (1.12)	1.16 (3.34)*
Party Frac.	0.55 (6.97)*	0.60 (6.07)*	0.78 (1.85)	1.01 (2.19)*	-0.21 (-0.87)	0.22 (0.90)
log(GDP p.c., PPP)	0.09 (4.39)*	0.09 (4.47)*	0.83 (7.32)*	0.84 (7.39)*	0.59 (7.68)*	0.60 (8.45)*
log(Population)	-0.03 (-3.42)*	-0.03 (-3.40)*	-0.17 (-3.21)*	-0.18 (-3.19)*	-0.25 (-7.21)*	-0.25 (-6.95)*
Urbanization	0.00 (0.31)	0.00 (0.12)	0.00 (0.74)	0.00 (0.62)	0.00 (-0.82)	0.00 (-1.24)
Other statistics						
Countries	135	135	135	135	135	135
Observations	224	224	224	224	224	224
Adjusted R-squared	0.51	0.51	0.55	0.55	0.56	0.60

Table 8: Elite-level ethnic cleavage and governance outcome. OLS with stand errors clustered at the group level. $*$ indicates $p < 0.05$; t-statistics in parentheses.

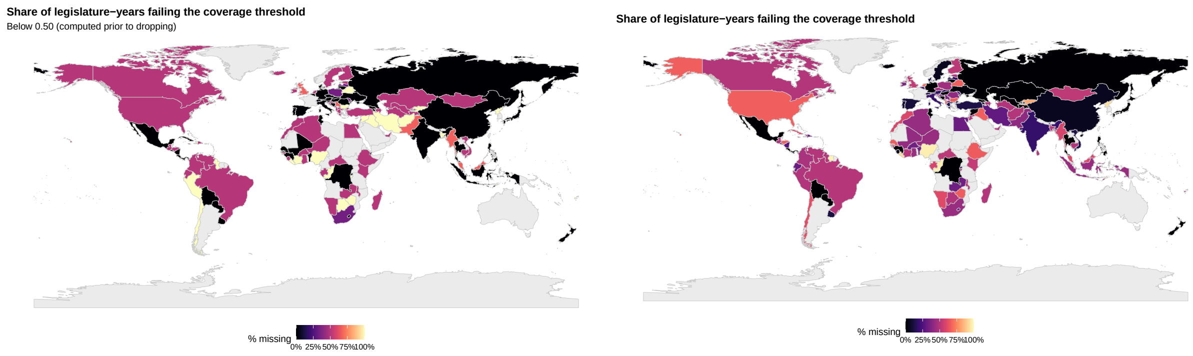
Share of legislature−years failing the coverage threshold
Below 0.50 (computed prior to dropping)
% missing
0%
25%
50%
75%
100%
Share of legislature−years failing the coverage threshold
% missing
0%
25%
50%
75%
100%

Figure 7: Illustrative sample selection. The maps show the share of legislature-years failing a minimum party-coverage threshold under expert-only coding (left) versus expert plus ASA augmentation with abstention (right).

Kaufman, Shachar, Saharon Rosset, and Claudia Perlich (2012): “Leakage in data mining: Formulation, detection, and avoidance”. In: *ACM Transactions on Knowledge Discovery from Data*, no. 4, vol. 6, 15:1–15:21. DOI: 10.1145/2382577.2382579.

Lewis, Patrick et al. (2020): “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *Advances in Neural Information Processing Systems*.

Montgomery, Jacob M, Brendan Nyhan, and Michelle Torres (2018): “How conditioning on posttreatment variables can ruin your experiment and what to do about it”. In: *American Journal of Political Science*, no. 3, vol. 62, pp. 760–775.

Paleka, Daniel et al. (2025): *Pitfalls in Evaluating Language Model Forecasters*. arXiv: 2506.00723. URL: https://arxiv.org/abs/2506.00723.

Pemstein, Daniel et al. (2020): *The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data (V-Dem Working Paper No. 21)*.

Peng, Roger D (2011): “Reproducible research in computational science”. In: *Science*, no. 6060, vol. 334, pp. 1226–1227. DOI: 10.1126/science.1213847.

Rasheed, Razeen A. et al. (2026): *From Fluent to Verifiable: Claim-Level Auditability for Deep Research Agents*. arXiv: 2602.13855. URL: https://arxiv.org/abs/2602.13855.

Stodden, Victoria, Peixuan Guo, and Zhaokun Ma (2013): "Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals". In: *PLOS ONE*, no. 6, vol. 8, e67111. DOI: `10.1371/journal.pone.0067111`.

Yao, Shunyu et al. (2022): *ReAct: Synergizing Reasoning and Acting in Language Models*. arXiv: `2210.03629 [cs.CL]`.